Sam Ogihara, Justin Chow
Professor Zimbra
OMIS 114
15 March 2022

Final Project Report

Playing and discussing video games has become a customary and important part of everyday life. Experts report that the "video game industry brought in a colossal $150 billion in revenue in 2020, and it's estimated that there are 2.7 billion gamers across the globe." (Field Level Media). Thus, we asked ourselves this question: What types of video games are people playing today? We attempted to answer that and other questions about video game sales, trends, and scores around the world. For our data science research project, our research question was: What are the best selling games in the modern history of gaming according to game sales and scores? With this question in mind, we examined the most sold games from a global perspective as our team analyzed the game's genre, platform, regional sales, and the year of release. In addition, our information about game publishers and platforms were obtained from public datasets. Diving deeper into our investigation of our research question, we had five sub-questions that we attempted to answer to assist us in our understanding of what we were researching. Those five questions were: How has the total sales of games changed over time? What does the video game landscape look like today? Are critic scores consistent with the sales of the best selling games? Do critics and regular gamers prefer different genres of games? What is the relationship between critic scores and global sales of games? All these questions are areas that our team concluded that were important in determining whether there is significant correlation between video game sales and people's preference of video games.

Our team's first dataset was selected from the publisher VGChartz (Video Game Charts): "a business intelligence and research firm … with an ever expanding game database with over 55,000 titles" (Walton, *VGChartz*). In our team's dataset, it contains a list of video games with game sales figures greater than 100,000 copies. Attributes to note are the 33 unique consoles, years of a game released, the genre, the 580 unique publishers, the regional sales (NA, EU, JP, and other regions) in the millions, and the total worldwide sales in the millions. When exploring the dataset, we observed that many of the existing games had their original and subsequent releases, which meant that franchise games were repeated because they had multiple releases on different consoles. In addition to this dataset, our team wanted to evaluate the performance of these games not only by their sales figures, but also by the gaming scores critics and regular gamers gave to games.

To incorporate the gaming scores to our exploration, our team's second dataset was selected from the Metacritic website: critics' consensus in one place with a single meta score and streamlining their user voting process (Editorial team, *Metacritic*). In our team's dataset, we were able to compile a list of critic and user scores of the best video games of all time. Attributes to note are the 22 unique consoles, release dates spanning 1996 to 2017, meta critic score (from 0 to 100) and user critic score (from 0 to 10). When our team observed the meta critic dataset, we noticed that there were much less observations. This can be explained by the 'All Platform' filter in place when the data was retrieved from the Metacritic website so each game includes scores from their original release (unless the games were part of a franchise). Though our datasets may contain a multitude of missing values, our team was confident in their ability to analyze the correlation between video game sales and scores.

Our team pulled the VGCharts and Metacritic datasets from the Kaggle website and imported them into the notebook. We proceeded to measure the length of specific columns in each dataset for unique data values and checked for NaN values. The columns of the second data frame were renamed to allow for an outer merge of two datasets - *vgsales.csv* and *all_games.csv* - calling the new dataframe *df_combined*. Then, our team created a new dataframe called *df_merged* grouping by name and aggregating mean row values of all numeric variables from *df_combined*. We checked the length of the combined datasets to see if the data values were filtered by the groupby function. Next, we cleaned *df_combined* by removing the rows with NaN values as some games were released in different years such as games being released on multiple consoles. Right after we cleared the NaN values, we filled NaN values for the Genre and Publisher categorical variables with the value 'Unknown' when the list of games was last updated. Using the dropna command drops any NaN value from rows that are mostly missing sales and scores values as a result of the outer merge. We cleaned the grouped dataframe *df_merged* and this left rows of game copies that are missing sales and score values. We applied the dropna function again to drop all the NaN values and rename the columns to match the *df_combined* dataframe. Once our team checked for NaN values using the *.isna().sum().sum()* command seeing that both *df_combined* and *df_merged* no longer contained NaN values, we started our exploration of video game sales and scores.

To better understand the data we would later visualize, we randomly sampled 500 rows from the *df_merged* dataframe. At this stage, our group focused on finding the correlation between sales figures and scores present in our analysis. We did this by defining a new dataframe called *dfv* that created a sample of 500 rows in a random state. With the *dfv* dataframe, we used the command *.describe()* to find the descriptive statistics of our sales and score variables. We were not surprised by the mean of the total worldwide sales being approximately equal to the addition of all the regional sales. However, the average number of total worldwide sales only reached around 1.5 million game copies sold. A poor sales average paired with a mean critic score below 70 and a user score below 7 suggests a strong correlation between global sales and scores. Our group decided to utilize a seaborn correlation heatmap to better see the impact of each correlation in our dataset. We found that the correlation between total worldwide sales of games and both scores were relatively weak at 0.34 for critics and 0.2 for regular gamers. The highest correlation our group found to 'Global Sales' was with NA sales (as expected); however, it is clear that all regional sales (including NA, EU, JP, and other regions) total to global sales so they are clearly correlated. Despite these findings, our team decided to further explore the merged dataset by applying visualization techniques to get a better understanding of the relationship between certain variables.

When observing the best selling games of all time, our group used the *.nlargest(3, columns = 'global_sales')* to find that Wii staple games published by Nintendo, on average, were the best selling games of all time. To better understand the yearly trend in global sales, our group decided to create a new dataframe from the merged dataset called *df_year* grouping by year and selecting all the regional sales columns to be summed into a total sales figure. When computing for the years with the largest number of games sold, our group determined that the mid-2000s was when global sales were highest, skyrocketing to around 370 million sales from 2006 to 2008. The majority of that change came from the exponential increase in NA sales in the same time period. To better understand the yearly trends in games sold, our team decided to graph a line plot using *plt.plot(df_year.Global_Sales, 'b')* where we observed the global sales of every game sold from *df_year*. We found that sales figures steadily rose from the 1990s into the 2000s, then it exponentially rose over 350 million sales in 2006 and again in 2008 before drastically falling.

Our group believes that the spike in yearly game sales, especially from 2006 to 2010, were heavily influenced by one of the top gaming publishers in our dataset. We decided to create a new dataframe from the combined dataset called *df_pub* grouping by publisher and computing the sum of global sales by each publisher in our dataset. When using the *.nlargest (5, columns = 'global_sales')* function*,* our group found that the top five publishers are Nintendo at around 728 million dollars, Electronic Arts at around 716 million dollars, then a steep drop for Activision at 447 million, Sony Computer Entertainment at 326 million, and Take-Two Interactive at 316 million. After we grouped by year and summed the yearly sales for all five gaming publishers, we used the *plt.plot* command to observe their yearly trend of game sales. Our group found that Nintendo's total worldwide sales of games after 2005 contributed to the spikes in sales in 2006 and 2008. Outside of Nintendo games, other top publishers had a steady rise in sales.

The success of consoles has had a major impact on publisher success so our team wanted to determine which console had sold the most games. We took a closer examination at the 'Platform' column in *df_combined* and we have found that the top three consoles with the most sold games are the Playstation 2 (853 million), Xbox 360 (665 million), and the Playstation 3 (578 million). Thus, the data implies that gamers ultimately prefer playing games on Playstation consoles over every other console. In addition, we used the catplot function to display a bar chart which confirms that the top platforms with the most sold games are non-Nintendo consoles; first the Playstation 2, followed by the Xbox 360, and the Playstation 3, with PC games not far behind in our dataset.

From our previous visualization of total games sold per console, our group found that the best-selling consoles were largely dominated by modern gaming systems including the PlayStation 2 (*PS2*), PlayStation 3 (*PS3*), and Xbox 360 (*X360*) having the most games released. We further wanted to find the top selling modern games of these platforms using the *df_combined* data frame specifying the value *PS2* for the platform column grouping by 'Name' to calculate the aggregate mean of global sales. When using the *.nlargest(10, columns='Global_Sales')* command, the Grand Theft Auto franchise of games before the PS3 remain the four highest sold games on the PS2 followed by Final Fantasy X. Our group conducted similar aggregate techniques for the PS3 finding that Grand Theft Auto V remained the most sold game, but saw more representation from the Call of Duty franchise of games. Finally, when evaluating the Xbox 360 top-selling games, we found it interesting that the best game sold was Kinect Adventures even though the Kinect was largely considered a failure by gamers and critics alike. Using visualization techniques helped us observe the true difference in sale figures between these top-selling consoles. When defining name and sales into separate variables then using the *plt.barh(name_'console', sales_'console')* command, we observed that the Xbox 360 had the highest selling games on average, which suggests overall better performance of the console despite it being ranked third in most games released.

Our group created a violin plot of the best-selling consoles against critic scores including Nintendo consoles Wii and DS. We found that the PS3 has the highest median score followed closely by the Xbox 360 and PS2. Each console shared a similar data distribution wider near their median score suggesting that the weight of critic scores are highly concentrated around the median value. Nintendo consoles followed a similar distribution but with a higher concentration of scores close to the median (reduces the frequency of scores far from the mean). On the other hand, the best-selling consoles against user scores are the PS2 followed by the Xbox 360, PS3, and Nintendo consoles. In these violin plots, the smaller scale of user scores for the list of games in our dataset led our data distributions to be more concentrated at the median value. The greater spread of critic scores on the top consoles held more weight than user scores, specifically the weight of scores on consoles.

In the *df_combined* dataframe, we utilized the groupby function to filter out the data by genre and the number of games sold globally. From the data table, it revealed that the genre of the game determines what regular gamers prefer based on the highest number of games sold. The data from the table was plotted on a bar graph, using the catplot function, and the *.nlargest(3, columns = 'Global_Sales')* command was applied. We came to the realization that in the top three video game genres that have the highest sales worldwide, the 'Action' genre had the largest total worldwide sales by far with over 1 billion games sold. This number was followed by the 'Sports' and 'Shooter' genre in our dataset.

To observe the yearly trend in the genre of games by critics and regular gamers, our group decided to visualize it in a categorical bar plot. Our team created a new dataframe called *df_genre* which used *df_combined* to group by genre and year, then computing the mean using the aggregate function to obtain the mean critic score for each genre. Our group also created four bins that each spanned an average of 4 to 6 years in order to see a more comprehensive visual. When evaluating the genre of games against user score, there was relatively no change occurring throughout the four binned years, suggesting a consistent criteria and preference of critics when it comes to scoring video games. On the other hand, the yearly trend in the genre of games by regular gamers tends to fluctuate much more in our categorical plot. Out of the top 3 genres, 'Sports' and 'Shooter' games decreased in average score suggesting that regular gamers have greater preference for specific genres as gaming interests change over time. Our group believes that as newer games shift to a more popular selling genre, their score for those games shift. In our plot, we can see the effect in the 'Adventure' and 'Racing' games as their higher average score steadily rises. Overall, our categorical bar plot reveals that regular gamers hold more opinions about a genre of a game as newer games are released, which could suggest the rise and demise of certain games more than what critics may suggest.

Our group created violin plots of the top genres according to critic scores, specifically, 'Action,' 'Sports,' and 'Shooter' games, to see the spread of the highest-rated genres by regular gamers. When evaluating our violin plots, 'Action' and 'Shooter' games have the highest user scores just above scores given to 'Sports' games. In these 3 violin plots, we see a similar data distribution that has extremely thin ends and is wider near the median. This suggests that the weight of user scores are highly concentrated around the median value. On the other hand, the highest-rated genres by critics, on average, are 'Sports' games. The greater spread of these 3 genres' data distribution means that critic scores given to a genre hold less weight and thus are less likely to draw closer to the median critic score.

Our team proceeded to observe critics and regular gamers preferences for consoles and genres. To answer our overarching research question, we brought our attention back to gamers' scores on sales figures. When examining the effect of metacritic scores on total worldwide sales, we see that critic scores and global sales have a clear positive correlation with each other. Also, as we evaluate critic scores on global sales in a subset of years, we can see that the slope, in the graphic, gets steeper past 80. Looking towards the effect of user score on total worldwide sales, we evaluated user scores on global sales in the whole dataset and we found a less positive correlation between the two variables. Then, when evaluating user scores in a subset of years, we discovered in the graphic that the slope slightly lowers past 6. Once a game gets a high user score, every additional point has a lower impact on global sales.

When developing a clustering analysis, our group's goal was to identify clusters of games published by Nintendo - has the most global sales of games - and their NA sales from our merged *df_combined* dataframe. First, we used the *pd.get_dummies(columns=['Publisher'], data=df)* to insert binary values for each row of games in our list. We also cleaned the transformed data frame by dropping the other non-numeric columns: *Name, Platform, Genre,* and *Summary*. Once we finished preparing the dataframe, we applied the k means clustering algorithm to identify 2 clusters. When we evaluated our generated summary info, we identified cluster 1 with more sold games in NA at 63.6 percent in NA with 6 percent of those games published by Nintendo than cluster 0 with only 21.9% of sold games with 3.2 percent of those games published by Nintendo. Investigate a finding of the clustering analysis, our group focused on the propensity of a Nintendo published game to be sold in NA. Nintendo published games only made about 10 percent of our list of games so we created a new column called Nintendo setting the condition *[df_combined].Publisher_Nintendo > 0.05* that separates the two clusters in our visuals.

In our clustering analysis, we chose to investigate Nintendo game sales in NA using visualization techniques such as categorical bar plots. In our first bar plot, we see that there is a strong natural tendency for their games to be sold in NA with an average of 1.5 million Nintendo game copies sold compared to less than 500,000 non-Nintendo game copies sold. To quantify the effect of Nintendo games over the years, we binned the years column once again to see that NA sales of the games on our list peaked in the first bin, but fell close to half a million in sales after 2000 and remained consistent since then. Ordering the plot with the categorical variable Nintendo, we re-evaluated the sales figure of Nintendo games and found that their games are selling, on average, much higher than non-Nintendo games in the NA region until 2012 peaking over 2 million sales.

Our group's first clustering analysis focused on NA sales of Nintendo games, but our group wanted to compare the NA video game market to other big markets, specifically Europe and Japan. Using similar visualization techniques from our first cluster analysis, we found that European sales of Nintendo game copies, on average, were less than 1 million while Japan sales of Nintendo games are even lower at less than 600,000 games copies. Both regions' total sales of Nintendo games being less than NA also reflect in our categorical bar plot where Europe's peak sales of Nintendo games only reached a difference, on average, around a million copies compared to non-Nintendo games. To better quantify our results, our team used the *df_combined* data frame to group by *Nintendo* and computing size and mean of NA, EU, and JP sales. Comparing the sales of all games in our list, the sample size is much smaller for published Nintendo games at 5.13 percent (=232 / 4519). The NA region remains the largest market for the foreign video game publisher Nintendo, even more than the European and domestic Japanese markets. Based upon silhouette score, we also determined that the best algorithm to use in our analysis was indeed 2 clusters with a silhouette score of 0.54. Because of this, our team believes that our mean clusters are well separated from each other and clearly distinguish the regional sales of Nintendo games.

Overall, we determined that scores and sales of video games are highly correlated and other factors associated with video games link back to its success. In our exploration, we were able to observe the different relationships to global sales of video games and in the end were able to conduct a clustering analysis that evaluated regional sales of video games from Nintendo, one of the largest video game publishers in the world. Our team recognizes that our analysis could be improved upon utilizing certain attributes such as scores clusters instead of regional sales of Nintendo games. In addition, we missed the opportunity to possibly incorporate a Twitch dataset filled with audience count and demographic data on video games streams, which contributes to the popularity of said video games. Despite this, we were still successful in examining the most sold games from a global perspective.

Works Cited

Ali, Arslan. “Sales of Video Games.” *Kaggle*, 8 Oct. 2020, https://www.kaggle.com/arslanali4343/sales-of-video-games

Walton, Brett. “VGChartz - Video game sales tracking, Video game journalism” *VGChartz website.* June 2005 http://www.vgchartz.com

Chavarria, Ignacio. “Predicting Video Game Hits with Machine Learning.” *Medium*, Towards Data Science, 2 Aug. 2017, https://towardsdatascience.com/predicting-hit-video-games-with-ml-1341bd9b86b0

Contractor, Deep. “Top Video Games 1995-2021 Metacritic.” *Kaggle*, 30 Jan. 2022, https://www.kaggle.com/deepcontractor/top-video-games-19952021-metacritic

Doyle, Marc; Doyle, Julie & Dietz, Jason. “Metacritic - Movie Reviews, TV Reviews, Game Reviews, and Music Reviews” *Paramount Streaming,* Metacritic website, Jan 2001, https://www.metacritic.com

Norris, Devin. “A Data Driven Exploration of Video Games - Sales and Scores.” *Medium*, Analytics Vidhya, 19 Feb. 2021, https://medium.com/analytics-vidhya/a-data-driven-exploration-of-video-games-sales-and-scores-3c77f1c6573c

Person. “Report: Gaming Revenue to Top $159B in 2020.” *Reuters*, Thomson Reuters, 12 May 2020, https://www.reuters.com/article/esports-business-gaming-revenues-idUSFLM8jkJMl